L2/13-128

ISO/IEC JTC1/SC2/WG2 **N4456**

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation

Doc type: Working Group Document
Title: Latvian and Marshallese Ad Hoc Report
Source: Latvian and Marshallese Ad Hoc
Author: Lisa Moore, IBM
Status: Ad Hoc Report
Action: For consideration by ISO/IEC JTC1/SC2/WG2
Date: 2013-06-11

An ad hoc on Latvian and Marshallese met in Vilnius, Lithuania on June 11, 2013. The following experts were present:

Deborah Anderson (USA)
Peter Constable (USA)
Michael Everson (Ireland)
Lisa Moore (USA)
Roberts Rozis (Latvia)
Andrew West (UK)
Ken Whistler (USA)

The meeting was chaired by Lisa Moore.

Documents:

The discussion in the ad hoc group was based on these current documents:

- N4449 – Cedillas and Commas Below
- N4450 – Latvian and Livonian Glyphs with Comma Accent in the Unicode Standard

Background:

In January 2013, the Unicode Technical Committee discussed issues for the representation of Marshallese orthography. In particular, Marshallese uses the Latin script and requires the letters l, m, n, and o with cedilla. Latvian orthography uses the Latin script and requires the letters g, k, l, n, and r with comma below. For Marshallese, it is unacceptable to display cedillas as commas below. Conversely, for Latvian, it is unacceptable to display commas below as cedillas.

Legacy practice has not been so precise. Latvian has historically used legacy ISO character sets that named the characters with cedilla, with the practice that rendering for Latvian would be done with a comma below. This ambiguity in legacy character sets was preserved when mapping to 10646/Unicode. The letters used by the Latvian language continued to use characters named with

cedilla, while having representative glyphs that used a comma below. Letters used by Livonian followed the same practice.

Due to this legacy practice that was extended into 10646 and Unicode, the Marshallese have found that the letters l and n with cedilla currently are consistently displayed with a comma below, contrary to their expectations for display.

Discussion:

The Latvian representative, Roberts Rozis, explained the history of character encoding in the Baltic states. The practice of naming Latvian characters with cedilla, but displaying them with a comma below began with the establishment of 8859-13. Also, Michael Everson pointed out that the first part of 8859 that included characters for Latvian was 8859-4.

The ad hoc discussed the different character encoding models used for different scripts in 10646/Unicode. Latin has a mixed model with some characters encoded with precomposed characters, some with decomposed characters, and some with both forms of representation.

Five possible ways of representing the l and n Marshallese letters were analyzed:

1. Do nothing. All members of the ad hoc agreed that this was not a good solution as it would only perpetuate the confusion.
2. Support ambiguous character representations. In this possible approach, using either comma or cedilla for the same abstract character would create further confusion and increase unexpected results for users.
3. Maintain Latvian (and Livonian) stability and use base letter plus combining cedilla for Marshallese. This approach violates the canonical equivalence of the two sequences.
4. Declare that the rendering of comma below and cedilla are always mutually distinct. With this approach, rendering one by the other would not be correct. This would, however, destabilize the current representation of the Latvian (and Livonian) languages.
5. Maintain Latvian (and Livonian) stability and encode four atomic characters with no decompositions, and with glyphs explicitly showing the cedilla form to support Marshallese:

   LATIN CAPITAL LETTER MARSHALLESE L WITH CEDILLA
   LATIN SMALL LETTER MARSHALLESE L WITH CEDILLA
   LATIN CAPITAL LETTER MARSHALLESE N WITH CEDILLA
   LATIN SMALL LETTER MARSHALLESE N WITH CEDILLA

Conclusion:

All members of the ad hoc agreed that option 5, encoding four new atomic characters to support Marshallese, would cause the least architectural disruption and would be the best way to proceed. Using this approach, stability would be preserved for the Latvian and Livonian user communities and also for the many implementations that currently support Latvian. It would also enable the correct representation of Marshallese with newly encoded, unambiguous forms for the required l and n with cedilla.

Various clarifying annotations on the characters used by Latvian and Livonian were also discussed, and it was agreed that any proposal for the four new Marshallese characters should include proposed annotations.